

How Many Instances Are Enough? Minimum Subsample Size Prediction for Stereotype Evaluation

Bora Bozdogan¹, Soroush Vosoughi², and Weicheng Ma³

^{1,2}Computer Science Department, Dartmouth College

³Department of Computer Science and Engineering, Oakland University

¹bora.bozdogan.28@dartmouth.edu

²soroush.vosoughi@dartmouth.edu

³weichengma@oakland.edu

Abstract

Stereotype evaluation for large language models relies on heavily annotated benchmarks, yet such evaluations are constrained by limited instance availability and are often sensitive to which examples are included. We propose a framework for predicting the minimum number of evaluation instances needed to preserve benchmark-level stereotype scores under specified error and confidence criteria. Our method provides a statistically grounded way to determine how much a benchmark can be reduced without materially changing the resulting evaluation outcome. We support the framework with theoretical analysis and extensive experiments on three widely used stereotype benchmarks, CrowS-Pairs, StereoSet, and BBQ. The framework assumes flattened, instance-level scoring rather than macro-averaging across bias categories. Under this formulation, the proposed method provides reliable subsample-size guidance in every dataset, model, and tolerance setting we test, with no undercoverage cases. As an additional extension, we show that our approach can also guide compact synthetic stereotype evaluation. Overall, our results suggest that benchmark-level stereotype evaluation can often be conducted with far fewer instances than the original datasets require, offering a practical and principled approach to more efficient bias assessment.

1 Introduction

Stereotype evaluation has become a standard component of responsible large language model (LLM) assessment, yet stereotype benchmarks are expensive to construct, and are inherently constrained by the number and distribution of available instances. In practice, this creates two related challenges. First, benchmark-based stereotype evaluation can be costly if similar conclusions could be reached from much smaller subsets. Second, static benchmarks are difficult to adapt to newly emerging bias

categories or dynamic evaluation settings, where generating and scoring large numbers of examples is often impractical.

In this paper, we address these challenges by asking a simple question: how many evaluation instances are actually needed to preserve a benchmark’s stereotype score within a desired error range? We propose a theoretical framework for predicting the minimum subsample size required to recover the full-dataset stereotype signal under specified tolerance and confidence criteria. Our approach is derived from a finite-population hypergeometric model, which makes the sample-size decision explicit and statistically interpretable. Rather than assuming that the full benchmark must always be used, the framework provides principled guidance on how many instances are needed to obtain nearly the same evaluation result.

We support this framework through both theoretical analysis and empirical validation. On three widely used stereotype benchmarks, CrowS-Pairs (Nangia et al., 2020), StereoSet (Nadeem et al., 2021), and BBQ (Parrish et al., 2022), the proposed method produces reliable subsample-size predictions across multiple model families and tolerance settings. At the benchmark level, it preserves full-dataset stereotype scores with high empirical fidelity while using substantially fewer instances than the original datasets, as low as 1% of the original dataset for some cases. We further show that the same pattern extends to LLM-synthesized stereotype evaluation data, indicating that the approach remains effective across different data sources.

This combination of accuracy and flexibility also makes the framework useful for dynamic stereotype evaluation. When benchmark-style instances must be generated for newly surfaced bias categories, constructing large synthetic datasets can be expensive and slow. We therefore use subsample-size prediction as a mechanism for guiding compact

synthetic benchmark construction. In a proof-of-concept framework, potentially biased categories are first surfaced dynamically, after which the required number of instances is determined by our method and only that many benchmark-style examples are generated and scored. This combines the flexibility of dynamic bias discovery with the rigor of benchmark-style stereotype measurement, while avoiding the cost of exhaustive dataset construction.

Our contributions are twofold. First, we introduce a framework for minimum subsample size prediction in stereotype evaluation and provide theoretical and empirical evidence that benchmark-level stereotype scores can be preserved with much smaller subsets on standard datasets. Second, we show that the same mechanism can guide compact synthetic stereotype evaluation, supporting more efficient and adaptive bias assessment.

2 Related Work

Bias evaluation in LLMs has been studied primarily through large-scale benchmark datasets such as StereoSet and CrowS-Pairs, which are expensive to construct and inherently static, limiting their ability to capture the broader and evolving bias behaviors of modern LLMs. Prior work has also documented conceptual weaknesses in existing bias benchmarks, including shallow representations of social groups and inconsistencies in how stereotyping is operationalized (Blodgett et al., 2021). More dynamic red-teaming and adversarial prompting approaches (Perez et al., 2022; Chao et al., 2023; Zhao et al., 2023) offer greater flexibility, but without the controlled structure of benchmark instances such as minimal pairs, results are harder to attribute to specific bias categories and more sensitive to prompt design.

Our work is most closely related to efforts that bridge this gap. One line of work studies benchmark subsampling, showing that relatively small subsets can preserve much of the original evaluation signal (Maia Polo et al., 2024; Perlitz et al., 2024). A second line uses LLMs to support data creation, including stereotype benchmark generation (Jha et al., 2023) and synthetic data generation more broadly (Wang et al., 2023; Honovich et al., 2023). Our work connects these two directions by proposing a statistically grounded framework for predicting how many benchmark instances are needed to preserve stereotype evaluation outcomes,

and showing how the same mechanism can guide compact synthetic stereotype evaluation.

3 Minimum Subsample Decision for Stereotype Evaluation

Prior work on stereotype evaluation typically assumes that reliable measurement requires using the full annotated benchmark, even though such benchmarks are costly to construct and often difficult to integrate into dynamic evaluation settings. At the same time, naive subsampling can be unstable because stereotype scores are sensitive to the distribution of evaluation instances, especially when the underlying bias rate is skewed or when the dataset is not large enough for asymptotic approximations to be reliable. We therefore propose a subsample computation procedure that explicitly models finite-population sampling and directly targets the probability that a subsample preserves the full-dataset stereotype score within a predefined tolerance.

Before stating the procedure, we fix the estimator it applies to. Throughout, the stereotype score is an instance-level binary mean over a flattened pool of evaluation instances, rather than a rate computed within each target group and then macro-averaged across groups, as StereoSet and common BBQ implementations do. Flattening does not discard the bias signal: both estimators target the same underlying quantity, namely the rate at which a model prefers stereotyped over anti-stereotyped options, and differ only in whether instances are weighted equally or reweighted by group size, so the instance-level mean remains a faithful benchmark-level measure of stereotyping. This choice is essential to the theory that follows. Under uniform sampling without replacement, the hypergeometric law is exact for a proportion of the form X/n , but a macro-average is an aggregate of ratios with random denominators: in a small subsample, group denominators become small and uneven, per-group rates grow highly variable and often extreme, and the variance of the aggregate far exceeds that of X/n . A subsample-size rule derived for X/n would therefore be optimistic if applied to a macro-averaged score. Flattening the aggregation level aligns the empirical estimator with the one the theory models, and marks precisely where additional structure would need to be modeled: for macro-averaged or otherwise random-denominator scores. With this convention in place,

we now derive the decision rule. Let a benchmark dataset contain N evaluation instances, of which K are labeled as biased. The full-dataset stereotype score is therefore $p_f = \frac{K}{N}$.

Suppose we draw a subsample of size n uniformly without replacement from this dataset, and let X denote the number of biased instances contained in the subsample. Since the sampling procedure is performed without replacement from a finite population, X follows a hypergeometric distribution: $X \sim \text{Hypergeometric}(N, K, n)$, with probability mass function

$$P(X=k) = \frac{\binom{K}{k} \binom{N-K}{n-k}}{\binom{N}{n}}.$$

The stereotype score computed on the subsample is then $p_s = \frac{X}{n}$.

We use the hypergeometric formulation rather than a normal approximation because it exactly reflects the finite-population setting and remains reliable even when the full-dataset stereotype score is near 0 or 1, where asymptotic approximations may become inaccurate. Although draws without replacement are not strictly independent, this dependence is accounted for by the hypergeometric model itself, which is precisely why it is appropriate in our setting.

Our goal is to determine the smallest subsample size n such that the subsample score remains sufficiently close to the full-dataset score with probability at least α . Let $\delta > 0$ denote the allowed absolute deviation. The target event is $|p_s - p_f| \leq \delta$.

Substituting $p_s = X/n$ and $p_f = K/N$, we obtain $|\frac{X}{n} - \frac{K}{N}| \leq \delta$. This is equivalent to

$$-\delta \leq \frac{X}{n} - \frac{K}{N} \leq \delta,$$

and after adding $\frac{K}{N}$ and multiplying through by n ,

$$n\left(\frac{K}{N} - \delta\right) \leq X \leq n\left(\frac{K}{N} + \delta\right).$$

Because X is integer-valued, the admissible range must be converted to integer bounds. Define

$$L_n = \lceil n\left(\frac{K}{N} - \delta\right) \rceil, \quad U_n = \lfloor n\left(\frac{K}{N} + \delta\right) \rfloor.$$

Then the probability that the subsample score falls within the target tolerance is exactly

$$P(|p_s - p_f| \leq \delta) = P(L_n \leq X \leq U_n),$$

where $X \sim \text{Hypergeometric}(N, K, n)$.

The minimum acceptable subsample size is therefore the smallest n for which this probability reaches the target confidence level α :

$$n^* = \min\{n \in \{1, \dots, N\} : P(L_n \leq X \leq U_n) \geq \alpha\}.$$

This is precisely the quantity computed by the proposed algorithm. Hence, under the finite-population sampling model, the algorithm returns the minimum subsample size whose stereotype score agrees with the full-dataset score up to deviation δ with probability at least α .

In our main experiments, the tolerance is specified relatively rather than absolutely. Specifically, for a relative tolerance parameter ε , we set $\delta = \varepsilon p_f$. Substituting this into the preceding derivation gives

$$L_n = \lceil np_f(1 - \varepsilon) \rceil, \quad U_n = \lfloor np_f(1 + \varepsilon) \rfloor,$$

and the minimum subsample size becomes

$$n^* = \min\{n \in \{1, \dots, N\} : P(L_n \leq X \leq U_n) \geq \alpha\}.$$

Therefore, the proposed decision rule is theoretically well founded: it is derived directly from the exact hypergeometric law governing subsampling without replacement, and it returns the smallest subsample size satisfying the desired tolerance and confidence requirements.

Overall, the proposed methodology addresses two limitations of prior practice. First, instead of assuming that the full benchmark must always be used, it provides a principled way to compute the smallest evaluation subset that preserves stereotype scores with high probability. Second, by relying on an exact finite-population model and an explicit empirical confirmation step, it avoids the instability of naive or asymptotic subsampling procedures and remains robust even in settings where stereotype rates are skewed or dataset composition materially affects the evaluation outcome.

3.1 Practical Use and Generality

Exact computation of n^* requires the full-dataset score p_f , so the framework applies whenever the score of a new evaluation can be anticipated from that of an earlier one. This condition is met in several common workflows. After one full run, the reduced set can be reused to score additional checkpoints, prompts, decoding settings, or interventions; a modified model can be sized using its parent model's score as an initial estimate; and new

bias categories that are structurally matched to existing ones can inherit their sizing. When no prior score is available, a pilot sample can be used to estimate p_f and the required size updated accordingly. In our own pipeline (Section 4.6), the same relationship is applied in the reverse direction, using the subsample score to predict the full-dataset score.

Although we frame this work around stereotype evaluation, nothing in the derivation is specific to stereotype scores. The only requirement is that the metric be an instance-level binary mean over a fixed denominator, so the method applies to any evaluation of this form. To confirm this, we validate the rule on a non-stereotype task, scoring 1,500 TweetEval stance instances with Qwen2.5-0.5B-Instruct, where the Bernoulli variable is classification correctness rather than stereotype preference. Across 10,000 trials per tolerance, the predicted sizes satisfy the uniform 95% criterion at $\varepsilon \in \{5\%, 10\%, 15\%, 20\%\}$, with the required sample decreasing from roughly 1,000 instances at $\varepsilon = 5\%$ to roughly 170 at $\varepsilon = 20\%$. We do not claim automatic transfer to benchmarks with macro-averaged scores, tie-dependent denominators, or other scoring structures, since those estimators would need to be modeled separately.

4 Empirical Verification of the Subsample Size Decision Algorithm

The theoretical analysis above specifies, for each dataset and tolerance level, a predicted minimum subsample size n^* that is expected to preserve the full-dataset stereotype score with coverage at least α . To assess whether this prediction is realized in practice, we conduct repeated subsampling experiments and estimate the empirical coverage achieved by n^* . Let $p_s^{(b)}$ denote the stereotype score obtained from the b -th subsample of size n^* , and define the indicator

$$I_b = \mathbf{1}\{|p_f - p_s^{(b)}| \leq \varepsilon p_f\},$$

which records whether the target tolerance criterion is satisfied in that trial. Over B independent subsampling trials, the empirical coverage is

$$\hat{\alpha} = \frac{1}{B} \sum_{b=1}^B I_b.$$

If α_t denotes the true coverage probability induced by the predicted subsample size, then $\sum_{b=1}^B I_b \sim \text{Binomial}(B, \alpha_t)$. For sufficiently large B , the

Central Limit Theorem yields

$$\hat{\alpha} \approx \mathcal{N}\left(\alpha_t, \frac{\alpha_t(1-\alpha_t)}{B}\right).$$

Under the target hypothesis $H_0 : \alpha_t = \alpha$, a one-sided 95% lower confidence bound, the confirmation threshold, is therefore

$$b_{\text{low}} = \alpha - 1.645 \sqrt{\frac{\alpha(1-\alpha)}{B}}.$$

We regard the predicted subsample size as empirically confirmed when $\hat{\alpha} \geq b_{\text{low}}$, meaning that the observed subsampling performance is statistically consistent with the target coverage level. Larger values of B make this confirmation stricter by narrowing the confidence bound. Based on this criterion, we conduct computational experiments in this section to empirically validate the proposed subsample size prediction algorithm. Specifically, we test whether the predicted minimum subsample sizes preserve full-dataset stereotype scores at the target coverage level across benchmark datasets and model settings.

4.1 Experimental Design

To instantiate the empirical validation framework, we use $B = 100,000$ subsampling trials throughout and consider four relative error tolerances, $\varepsilon \in \{0.05, 0.10, 0.15, 0.20\}$. Predicted subsample sizes n^* are computed at a coverage level of 0.95 throughout, and for empirical confirmation we adopt a target level of 0.95 which corresponds to the b_{low} thresholds reported in each table. Our experiments cover three stereotype benchmarks, namely BBQ, StereoSet, and CrowS-Pairs. These datasets differ in construction and scoring format, allowing us to test the proposed subsample-size prediction method across multiple benchmark styles and bias measurement settings. We use the full datasets for StereoSet and CrowS-Pairs, and a random 6,591-instance subset of BBQ to keep computation feasible.

Seven models are included in the experiments, namely claude-sonnet-4-6 (Claude-Sonnet), gemini-2.5-flash (Gemini-Flash), gpt-4o (GPT-4o), gpt-5 (GPT-5), LLaMa-3.1-8B-Instruct (LLaMa-8B), Ministral-3-14B-Instruct-2512 (Ministral-14B), and Qwen2.5-7B-Instruct (Qwen-7B). We use HuggingFace checkpoints for the open-source models and the official APIs for the closed-source models. Because these models expose different interfaces, we use three scoring schemes. For GPT-5, Gemini-Flash, and Claude-Sonnet, which do not

expose token-level probabilities, we use a forced-choice setup in which each instance is presented as a multiple-choice question and the selected option determines the score. For LLaMa-8B, Ministral-14B, and Qwen-7B, we use direct sentence log probabilities and select the sentence with the higher probability. For GPT-4o, we use a top-logprob variant in which we aggregate the top 20 token-level log probabilities for each sentence and select the option with the higher resulting score.

Across benchmarks, ties are handled conservatively. In StereoSet, tied log-probability scores are treated as anti-stereotype. In CrowS-Pairs pseudo-probability mode, ties form a separate bucket and are excluded from the score. In BBQ, ties are broken by index order, although exact equality is practically negligible. We do not apply length normalization, since paired options are already closely matched in length: StereoSet pairs are identical by construction, CrowS-Pairs differs only minimally because of its minimal-pair design, and in BBQ the answer options are short phrases conditioned on much longer contexts, making small answer-length differences negligible.

4.2 Results and analyses

Figure 1 provides an overall view of how the proposed method behaves on BBQ, CrowS-Pairs, and StereoSet at a tolerance of 10%, while results at other tolerances and the detailed numerical values are reported in Appendix B. Across all three benchmarks, the histograms are centered close to zero and the cumulative curves show that most subsamples remain within the target error bands, indicating that the predicted subsample sizes preserve the full-dataset stereotype scores.

This visual pattern is reflected in the validation results. Overall, 84 of 84 model $\times$ dataset $\times$ tolerance cells are confirmed, corresponding to a pass rate of 100%. More broadly, the results are highly stable across datasets and tolerances. All three benchmarks are confirmed in every reported configuration, including all 21 cells at the strictest setting $\varepsilon = 5\%$. At the model level, Claude-Sonnet, Gemini-Flash, GPT-4o, GPT-5, LLaMa-8B, Ministral-14B, and Qwen-7B pass every configuration.

The figures also suggest that the method is broadly consistent across model families. For all benchmarks, the cumulative distribution function (CDF) and coverage curves are tightly aligned across models, and the histograms remain narrow

and centered. These patterns suggest that the proposed method is not tied to any one model family or benchmark format.

The subsample-size predictions themselves follow a clear and intuitive pattern. As the tolerance is relaxed, the required sample size decreases substantially for all three benchmarks. Averaged across models, BBQ requires 30.8% of the full dataset at $\varepsilon = 5\%$, dropping to 2.8% at $\varepsilon = 20\%$. StereoSet is the most subsampling-efficient, requiring 13.5% at $\varepsilon = 5\%$ and approximately 1% at $\varepsilon = 20\%$. CrowS-Pairs requires the largest proportion at tighter tolerances, averaging 41.9% at $\varepsilon = 5\%$, which is expected given its smaller total size ($N = 1506$).

These results show that the proposed minimum subsample size prediction method provides reliable guidance on how many evaluation instances are needed to preserve benchmark-level stereotype scores on three widely used stereotype datasets. Rather than requiring the full benchmark by default, the method can identify much smaller subsample sizes that still maintain the resulting evaluation conclusions with high fidelity.

4.3 Synthetic Instance Construction

To extend the evaluation beyond the original benchmarks, we construct synthetic stereotype evaluation instances by first prompting a generator LLM to produce stereotype triplets of the form (g_s, g_a, r) , where g_s denotes a stereotyped social group, g_a denotes its anti-stereotyped counterpart, and r denotes the associated stereotype rationale. For example, a generated triplet may take the form (old people, young people, difficulties with technology). These triplets are then instantiated into the input formats required by each benchmark, so that the resulting synthetic instances are directly comparable to their real-data counterparts under the same evaluation protocol. Appendix A compares the synthetic and real instances and reports a small-scale manual validation, confirming the quality of the generated data points and their improvements over the original benchmark instances.

4.4 Results and analyses

Figure 2 summarizes the stereotype evaluation results on synthetic data at a 10% tolerance, while the remaining figures at other tolerances and detailed numerical values are reported in Appendix C. Across BBQ, CrowS-Pairs, and StereoSet, the deviation distributions are centered close to zero, and

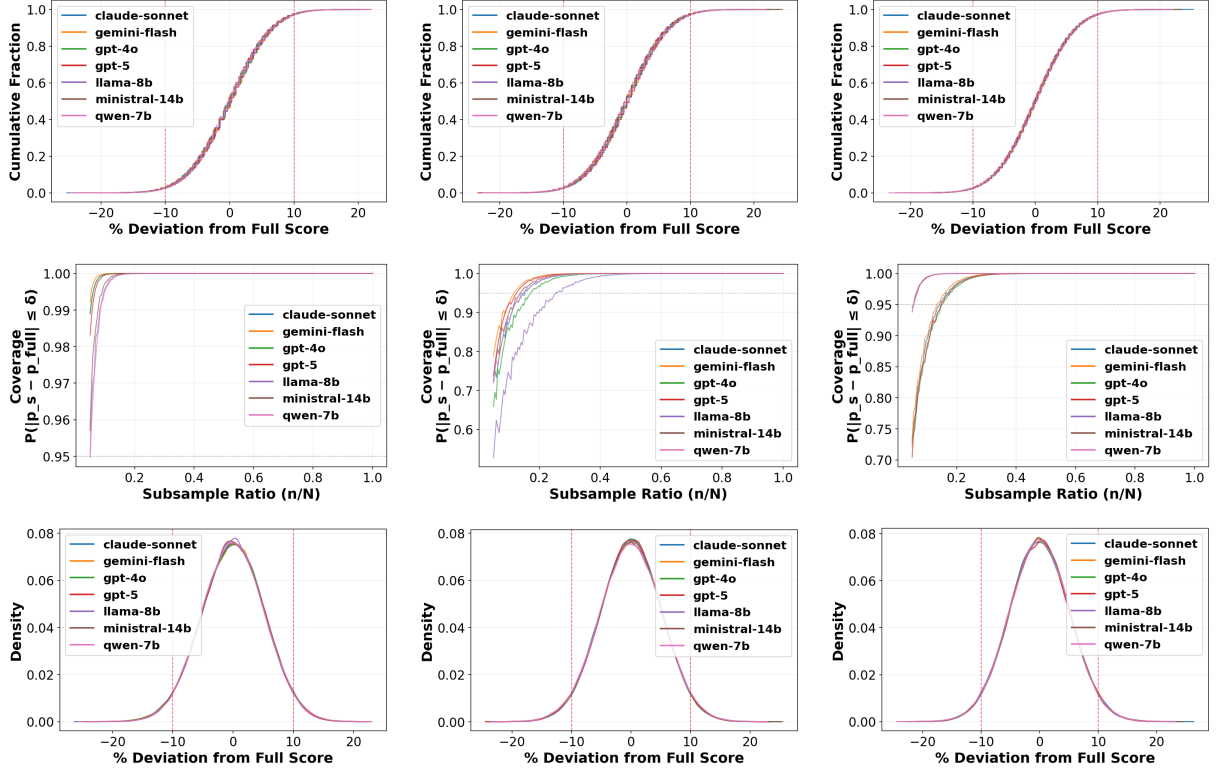


Figure 1: Subsampling behavior on StereoSet, CrowS-Pairs, and BBQ (left to right) at tolerance $\varepsilon=10\%$. Top: empirical cumulative distribution of percent deviation from the full-dataset score. Middle: empirical coverage versus subsample ratio. Bottom: kernel density estimation (KDE) of percent deviation from the full-dataset score.

the cumulative curves show that most subsamples remain within the target error band, suggesting that the predicted subsample sizes still preserve the benchmark-level stereotype signal effectively.

Overall, 84 of 84 model $\times$ dataset $\times$ tolerance cells are confirmed, corresponding to a pass rate of 100%. BBQ, CrowS-Pairs, and StereoSet are all confirmed in every reported configuration, and no failures occur on any dataset. At the model level, Claude-Sonnet, Gemini-Flash, GPT-4o, GPT-5, LLaMa-8B, Ministral-14B, and Qwen-7B pass all configurations. Confirmation holds even in the strictest settings, indicating that the synthetic-data results are strong overall and that no residual mismatches remain.

The results also highlight clear differences in required subsample ratios across synthetic benchmarks. On synthetic BBQ, the coverage curves rise more gradually and start from lower ratios than on the other two datasets, indicating that larger subsamples are needed before the target coverage is reached. This is reflected in the predicted sample sizes: BBQ requires 72.0% of the full synthetic dataset on average at $\varepsilon = 5\%$, dropping to 22.1% at $\varepsilon = 20\%$. This pattern is largely driven by several

low synthetic p_f values, which make tight relative tolerances especially demanding. By contrast, synthetic CrowS-Pairs and StereoSet are considerably more efficient, both requiring roughly one quarter of the dataset at $\varepsilon = 5\%$ and only about two percent by $\varepsilon = 20\%$.

Compared with the results on existing benchmarks, the synthetic-data evaluation shows the same broad behavior. In both settings, deviations remain centered near zero, model-level trends are generally stable, and the method achieves full validation rates with no failures. The main difference lies in the required subsample ratios rather than in validation outcomes. On existing benchmarks, the largest ratios occur on CrowS-Pairs, whereas on synthetic data they occur on BBQ. Even so, the overall validation rate on synthetic data, 100%, matches the 100% obtained on existing benchmarks, indicating that the method transfers well from benchmark reduction to synthetic benchmark construction.

More broadly, the synthetic-data results reinforce the earlier conclusions. The proposed subsample size decision algorithm remains effective when applied to generated stereotype benchmarks,

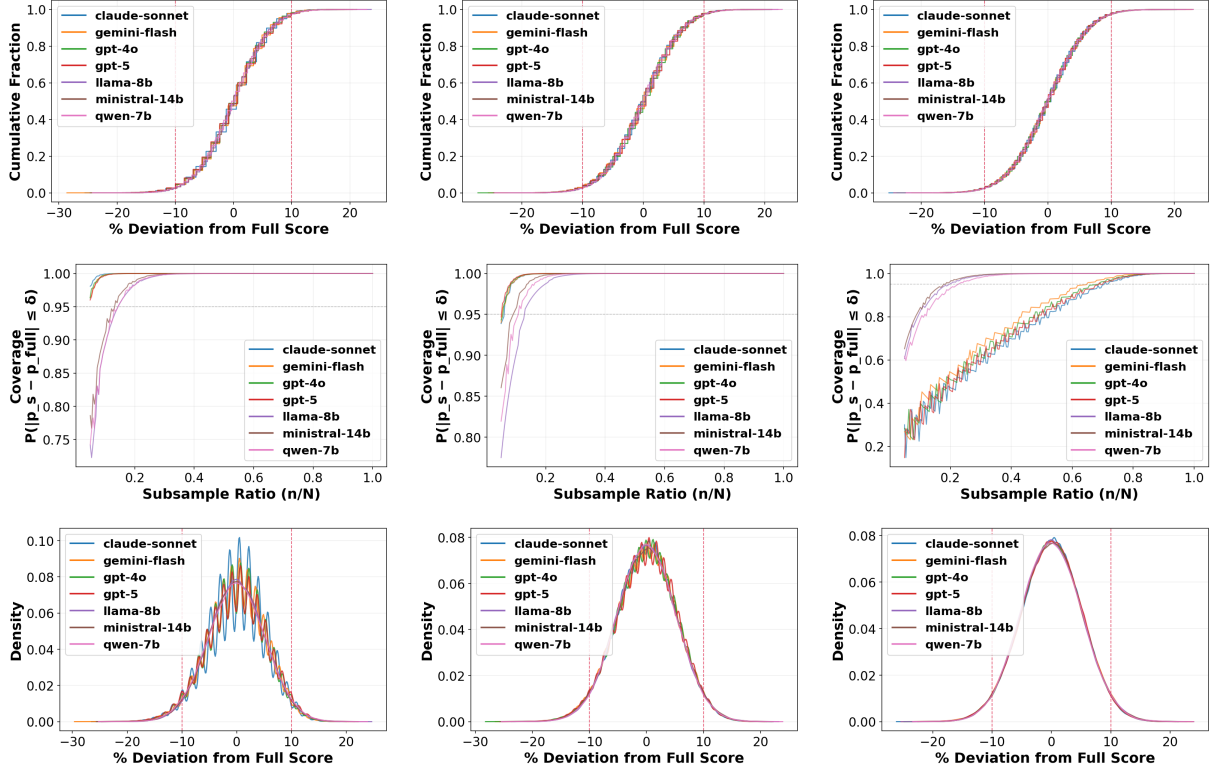


Figure 2: Subsampling behavior on synthetic data generated in the formats of StereoSet, CrowS-Pairs, and BBQ (left to right) at tolerance $\varepsilon=10\%$. Figure ordering is the same as in Figure 1.

and its qualitative behavior remains close to that observed on existing benchmarks. At the same time, the synthetic setting shares the same basic property identified earlier: under the strictest tolerances, the required subsample size grows substantially when the full-dataset stereotype rate is low, since the absolute tolerance εp_f becomes correspondingly narrow. This effect remains limited in scope and does not compromise confirmation at any tolerance level.

4.5 Category-Specific Subsample Size Prediction Results

We further extend the analysis from dataset-level scores to specific bias subcategories. Table 1 reports the category-level validation results for the proposed subsample size prediction method on selected benchmark subsets. For each dataset, four configurations spanning different model, dataset type, and bias category combinations are selected at random, providing a representative cross-section of the finer-grained evaluation setting. The category-level results suggest that the proposed subsample-size prediction behaves similarly under finer-grained subset evaluation, showing that the proposed method generalizes to category-specific

evaluations with the same reliability as the full-dataset setting. All BBQ, CrowS-Pairs, and StereoSet configurations clearly meet the 95% confirmation threshold across all tested categories and tolerances. What changes at the category level is not whether the predictions are confirmed but how large the required subsets become: because each category subset is far smaller than the full benchmark, the predicted sizes occupy a much larger share of the available instances, averaging 88.0% on BBQ, 77.4% on CrowS-Pairs, and 43.4% on StereoSet at $\varepsilon = 5\%$. The effect is most pronounced when the category-level stereotype rate is low, as in the synthetic BBQ race/ethnicity subset ($p_f = 1.3\%$), where the narrow absolute tolerance requires essentially the entire subset regardless of ε . StereoSet remains the most subsampling-efficient at this granularity, dropping to 6.3% of the subset by $\varepsilon = 20\%$, while BBQ still requires 46.5%. However, with only 12 total configurations across the three datasets, these comparisons are not conclusive, and the observed differences across datasets should be treated as indicative rather than definitive. Overall, the category-level results reinforce rather than contradict the full-dataset findings, suggesting that the proposed method is broadly applicable

ε	Category: p_f : N :	BBQ				CrowS-Pairs				StereoSet			
		Claude-Sonnet (Synthetic) Race/Eth.	GPT-5 Race/Eth.	Min.14B (Synthetic) Gender Id.	Qwen-7B Gender Id.	Gemini-Flash Race/Col.	GPT-5 (Synthetic) Race	LLaMa-8B Gender	Qwen-7B (Synthetic) Gender	Claude-Sonnet Race	GPT-4o (Synthetic) Race	LLaMa-8B (Synthetic) Prof.	Min.14B Prof.
		1.3% 224	22.8% 597	37.9% 224	48.4% 599	58.6% 514	61.8% 228	53.8% 262	60.5% 233	57.4% 2976	85.8% 500	46.0% 500	55.2% 2398
5%	n^* (%)	221 (98.7%)	531 (88.9%)	205 (91.5%)	437 (73.0%)	345 (67.1%)	182 (79.8%)	217 (82.8%)	186 (79.8%)	821 (27.6%)	168 (33.6%)	392 (78.4%)	820 (34.2%)
	Δpp	0.07	1.14	1.90	2.42	2.93	3.09	2.69	3.03	2.87	4.29	2.30	2.76
	$\hat{\alpha}$	0.960	0.958	0.953	0.957	0.955	0.959	0.950	0.957	0.955	0.959	0.949	0.954
	b_{low}	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949
	n^* (%)	221 (98.7%)	404 (67.7%)	161 (71.9%)	243 (40.6%)	174 (33.9%)	118 (51.8%)	142 (54.2%)	119 (51.1%)	255 (8.6%)	49 (9.8%)	234 (46.8%)	277 (11.6%)
10%	Δpp	0.13	2.28	3.79	4.84	5.86	6.18	5.38	6.05	5.74	8.58	4.60	5.52
	$\hat{\alpha}$	0.960	0.953	0.954	0.955	0.953	0.959	0.954	0.956	0.954	0.952	0.953	0.954
	b_{low}	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949
	n^* (%)	221 (98.7%)	294 (49.2%)	124 (55.4%)	133 (22.2%)	94 (18.3%)	72 (31.6%)	89 (34.0%)	75 (32.2%)	120 (4.0%)	26 (5.2%)	138 (27.6%)	130 (5.4%)
	Δpp	0.20	3.42	5.69	7.26	8.78	9.28	8.07	9.08	8.61	12.87	6.90	8.28
15%	$\hat{\alpha}$	0.960	0.960	0.962	0.952	0.951	0.960	0.951	0.955	0.952	0.963	0.956	0.955
	b_{low}	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949
	n^* (%)	221 (98.7%)	202 (33.8%)	88 (39.3%)	85 (14.2%)	57 (11.1%)	46 (20.2%)	62 (23.7%)	47 (20.2%)	67 (2.3%)	10 (2.0%)	89 (17.8%)	76 (3.2%)
	Δpp	0.27	4.56	7.59	9.68	11.71	12.37	10.76	12.10	11.49	17.16	9.20	11.03
	$\hat{\alpha}$	0.960	0.951	0.951	0.954	0.955	0.960	0.959	0.955	0.955	0.959	0.954	0.954
20%	b_{low}	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949

Table 1: Minimum subsample prediction on bias category subsets. Each column corresponds to a specific benchmark/model/condition/category combination. Rows report, for each relative tolerance ε , the required subsample size n^* (% of N), the allowed deviation in percentage points ($\Delta pp = \varepsilon \cdot p_f$), the empirical coverage $\hat{\alpha}$, and the confirmation lower bound b_{low} . The full-set stereotype score p_f and subset size N are shown in the column header.

to finer-grained bias analyses, provided that the reduced subset size is taken into account when interpreting the resulting savings.

4.6 Subsample Size-Guided Data Synthesis

The preceding analysis motivates a practical extension from static benchmark reduction to dynamic benchmark construction. Once a potentially biased category has been identified, the earlier subsample-size results can be used to guide how many synthetic instances need to be generated in order to obtain a stable benchmark-style bias estimate, without incurring the cost of building a large static dataset. In this way, the proposed framework uses the subsample-size decision mechanism not only to compress existing benchmarks, but also to inform the creation of new evaluation sets for emerging bias categories.

Concretely, we develop a dynamic synthetic evaluation framework that combines adaptive bias discovery with benchmark-style measurement. The framework proceeds in three stages. First, candidate demographic categories are screened through a taxonomist–actor–critic process, in which a taxonomist LLM proposes categories, an actor LLM converts them into probing prompts, and a critic LLM determines whether the target model’s response exhibits biased or stereotyped framing. Second, only categories flagged as biased are expanded into synthetic benchmark instances in a standard evaluation format, such as CrowS-Pairs or StereoSet. Third, the target model is evaluated on these generated instances using the same scoring procedures applied to the original benchmarks. This

pipeline allows dynamically discovered bias categories to be converted into compact, benchmark-formatted evaluation sets whose size is informed by the proposed subsample-size analysis.

4.7 Synthetic Stereotype Evaluation Proof-of-Concept Validation

We illustrate the framework with a single end-to-end run in which Claude-Sonnet serves as the evaluator and Gemini-Flash serves as the target model for evaluation. In this run, the generated instances use the CrowS-Pairs format. During the category-detection phase, 15 demographic categories are examined. Of these, three are flagged as biased while the remaining 12 are judged non-biased. The biased groups are *bisexual people*, *elderly women*, and *refugees*.

Only the three flagged categories proceed to benchmark construction. For each category, the evaluator generates 50 CrowS-Pairs instances, yielding 150 synthetic benchmark items in total. Each instance consists of a more-stereotyped sentence and a less-stereotyped counterpart that differ only in the demographic group referenced. The target model is then evaluated by forced choice on each pair, and the resulting CrowS score is computed as the proportion of stereotyped choices, with 50 corresponding to a neutral baseline.

The resulting scores are all above the neutral threshold, consistent with the critic’s earlier judgments. *Elderly women* shows the clearest effect, with a CrowS score of 86.0 (43 of 50 stereotyped choices), followed by *refugees* at 60.0 (30 of 50) and *bisexual people* at 54.0 (27 of 50). No ties oc-

cur in any category. These results provide a proof of concept that a relatively small synthetic benchmark, whose size is informed by the earlier subsample analysis, can recover a clear downstream bias signal for categories surfaced during the dynamic discovery stage.

As an additional demonstration, we run a small baseline-style comparison in the StereoSet format. We generate 40 synthetic instances for each of four categories: two that the critic identifies as biased and two that it identifies as non-biased. The biased categories score above the neutral stereotype baseline (*rural communities*: 69.9; *teenage mothers*: 57.5), whereas the non-biased categories score below it (*polyamorous individuals*: 40.2; *gender people*: 32.4), and the associated bias rates follow the same ordering.

Both demonstrations are intended as proofs of concept rather than as evaluation results in their own right. With a single end-to-end run and four categories in the comparison above, the evidence is too limited to support claims about the reliability of the critic or about the bias behavior of any particular model. What the runs do show is that the pipeline is executable end to end: dynamically surfaced categories can be converted into benchmark-formatted instances whose size is informed by the subsample-size analysis, and those instances yield stereotype scores that are ordered consistently with the critic’s initial judgments. Constructing and validating a full dynamic benchmark, which would require systematic category coverage, human annotation of the generated instances, and validation across target models, is a substantially larger undertaking and is out of scope for this paper.

Taken together, these examples show how the proposed subsample-size analysis can be operationalized in a dynamic setting. Rather than requiring a large pre-existing benchmark, the framework identifies potentially biased categories, generates a compact number of benchmark-formatted instances, and applies standard stereotype scoring to quantify the resulting bias signal, demonstrating the practical value of the method beyond static benchmark reduction alone.

5 Conclusion and Future Work

This paper presented a minimum subset prediction framework for estimating the minimum subsample size needed to preserve stereotype evaluation outcomes under target error and confidence crite-

ria. Experiments on BBQ, CrowS-Pairs, and StereoSet show that the proposed method can recover benchmark-level stereotype signals with substantially fewer instances while remaining robust across models and evaluation settings. We also demonstrated that the same idea can support dynamic synthetic-data construction, providing a practical bridge between static benchmark evaluation and flexible bias discovery.

Future work should validate the framework on a broader range of synthetic-data generation pipelines and extend it to finer-grained bias analyses, including subgroup and intersectional settings.

Limitations

This work has a important limitations that we do not attempt to fully resolve here. Our synthetic-data experiments are intended as a proof of concept showing that the proposed approach can effectively augment synthetic stereotype evaluation by providing principled guidance on how many instances to generate. In other words, the contribution here is not to exhaustively compare all possible synthesis pipelines, but to show that subsample size prediction can serve as a practical mechanism for making synthetic evaluation more compact and statistically grounded. A full-scale evaluation across different generation frameworks, prompting strategies, and evaluator designs would be costly and broad enough to merit a dedicated study of its own.

Additionally, we focus on preserving aggregate stereotype scores, rather than optimizing for every possible downstream use of a benchmark. In practice, some applications may require stronger guarantees for subgroup-level estimates, intersectional slices, or rare bias categories. These settings are important, but they also introduce smaller effective sample sizes and more fragile statistical behavior, which would require additional theory and validation. This does not weaken our central findings on benchmark-level stereotype evaluation, but it does define a natural and important direction for future work.

Ethics Statement

This paper studies stereotype evaluation, a topic that necessarily involves harmful social content. However, our goal is not to generate, amplify, or disseminate stereotypes. Instead, we focus on improving the efficiency and flexibility of methods for detecting and mitigating stereotypical behavior

in language models. In this sense, the contribution of the paper is protective rather than generative: we aim to make stereotype evaluation more statistically grounded and more practical to deploy.

The synthetic-data component should be understood in the same way. The generated instances are used only as controlled evaluation artifacts for stereotype detection and mitigation, not as content to be shared, deployed, or reused outside the experimental setting. We do not release these generated stereotype instances as a public resource, and we do not present the framework as a means of producing or spreading harmful content. Because the synthesis experiments are tightly scoped to benchmark-style evaluation, we do not foresee additional ethical risks beyond those already inherent in standard bias-evaluation research.

We also used large language models to assist with limited aspects of manuscript preparation, specifically for light writing support and polishing. The core scientific contributions of the paper, including the problem formulation, methodology, theoretical development, experiments, analysis, and conclusions, were developed by the authors rather than delegated to language models.

References

- Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. [Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1004–1015, Online. Association for Computational Linguistics.
- Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J. Pappas, and Eric Wong. 2023. Jailbreaking black box large language models in twenty queries. *arXiv preprint arXiv:2310.08419*.
- Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. 2023. [Unnatural instructions: Tuning language models with \(almost\) no human labor](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14409–14428, Toronto, Canada. Association for Computational Linguistics.
- Akshita Jha, Aida Mostafazadeh Davani, Chandan K. Reddy, Shachi Dave, Vinodkumar Prabhakaran, and Sunipa Dev. 2023. [SeeGULL: A stereotype benchmark with broad geo-cultural coverage leveraging generative models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9851–9870, Toronto, Canada. Association for Computational Linguistics.
- Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. 2024. [tinyBenchmarks: Evaluating LLMs with fewer examples](#). In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, pages 34303–34326. PMLR.
- Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. [StereoSet: Measuring stereotypical bias in pretrained language models](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.
- Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. [CrowS-pairs: A challenge dataset for measuring social biases in masked language models](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.
- Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R. Bowman. 2022. [BBQ: A hand-built bias benchmark for question answering](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105, Dublin, Ireland. Association for Computational Linguistics.
- Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. [Red teaming language models with language models](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3419–3448, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Yotam Perlitz, Elron Bandel, Ariel Gera, Ofir Arviv, Liat Ein-Dor, Eyal Shnarch, Noam Slonim, Michal Shmueli-Scheuer, and Leshem Choshen. 2024. [Efficient benchmarking \(of language models\)](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Mexico City, Mexico. Association for Computational Linguistics.
- Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. [Self-instruct: Aligning language models with self-generated instructions](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.
- Jiaxu Zhao, Meng Fang, Shirui Pan, Wenpeng Yin, and Mykola Pechenizkiy. 2023. [GPTBIAS: A compre-](#)

hensive framework for evaluating bias in large language models. *arXiv preprint arXiv:2312.06315*.

A Synthetic Data Validation

Table A1: Linguistic quality comparison for BBQ (Real: 6,591 | Synthetic: 2,472).

	Real	Synthetic
<i>Sentence length (words)</i>		
Mean	43.65	42.99
Std Dev	22.67	16.62
<i>Dependency tree depth</i>		
Mean	5.28	5.72
Std Dev	1.76	1.61
<i>CoLA acceptability</i>		
Acceptable	98.8%	100.0%
Unacceptable	1.2%	0.0%
<i>Vocabulary (JSD)</i>		
JSD distance	0.5491	
JSD divergence	0.3015 bits	
Vocab size	2,812	6,364
Union vocab	7,364	

Table A3: Linguistic quality comparison for StereoSet (Real: 6,392 | Synthetic: 2,000).

	Real	Synthetic
<i>Sentence length (words)</i>		
Mean	8.88	10.93
Std Dev	3.30	1.68
<i>Dependency tree depth</i>		
Mean	3.10	3.83
Std Dev	1.15	1.09
<i>CoLA acceptability</i>		
Acceptable	83.1%	96.2%
Unacceptable	16.9%	3.8%
<i>Vocabulary (JSD)</i>		
JSD distance	0.6432	
JSD divergence	0.4136 bits	
Vocab size	7,539	3,934
Union vocab	9,361	

We validate the synthetic stereotype benchmarking instances by comparing them against the original datasets used to generate these data points along both surface-level and embedding-level dimensions. Tables A1–A3 summarize the lexical and syntactic comparisons for BBQ, CrowS-Pairs, and StereoSet, respectively. In addition to these tabular statistics, we analyze sentence embedding distributions using t-SNE projections over SBERT and BGE-Large-en-v1.5 embeddings.

Overall, the synthetic instances are comparable to the original data across all three benchmarks. Sentence-length means are close for BBQ (43.65 for real data vs. 42.99 for synthetic data) and reasonably similar for StereoSet (8.88 vs. 10.93),

Table A2: Linguistic quality comparison for CrowS-Pairs (Real: 1,506 | Synthetic: 2,083).

	Real	Synthetic
<i>Sentence length (words)</i>		
Mean	14.71	11.03
Std Dev	6.08	2.26
<i>Dependency tree depth</i>		
Mean	4.10	4.04
Std Dev	1.45	1.18
<i>CoLA acceptability</i>		
Acceptable	94.4%	99.6%
Unacceptable	5.6%	0.4%
<i>Vocabulary (JSD)</i>		
JSD distance	0.7764	
JSD divergence	0.6028 bits	
Vocab size	3,988	3,540
Union vocab	6,535	

while CrowS-Pairs shows a larger difference (14.71 vs. 11.03). Across all datasets, the synthetic data exhibits lower standard deviations in sentence length, indicating more uniform generation. Dependency tree depth is also highly similar between real and synthetic instances, suggesting comparable syntactic complexity.

The synthetic data achieves higher grammatical acceptability than the original data on all three benchmarks. This improvement is especially pronounced for StereoSet, where CoLA acceptability rises from 83.1% on the original data to 96.2% on the synthetic data. BBQ is already close to ceiling on the original benchmark (98.8%), while the synthetic data reaches 100%. These results suggest that the generated instances are at least as grammatically well formed as the original benchmark examples.

Vocabulary statistics show moderate divergence between real and synthetic datasets. Jensen–Shannon divergence is lowest for BBQ (0.3015 bits), intermediate for StereoSet (0.4136 bits), and highest for CrowS-Pairs (0.6028 bits). The synthetic data also tends to use a somewhat more constrained and formal vocabulary, whereas the original benchmarks more often contain casual or conversational wording. At the same time, these differences should be interpreted in light of dataset size: for example, the original StereoSet contains 6,392 instances and 7,539 unique words, whereas its synthetic counterpart contains 2,000 instances and 3,934 unique words.

Additionally, the t-SNE projections (Figure A1) show that the synthetic instances remain broadly

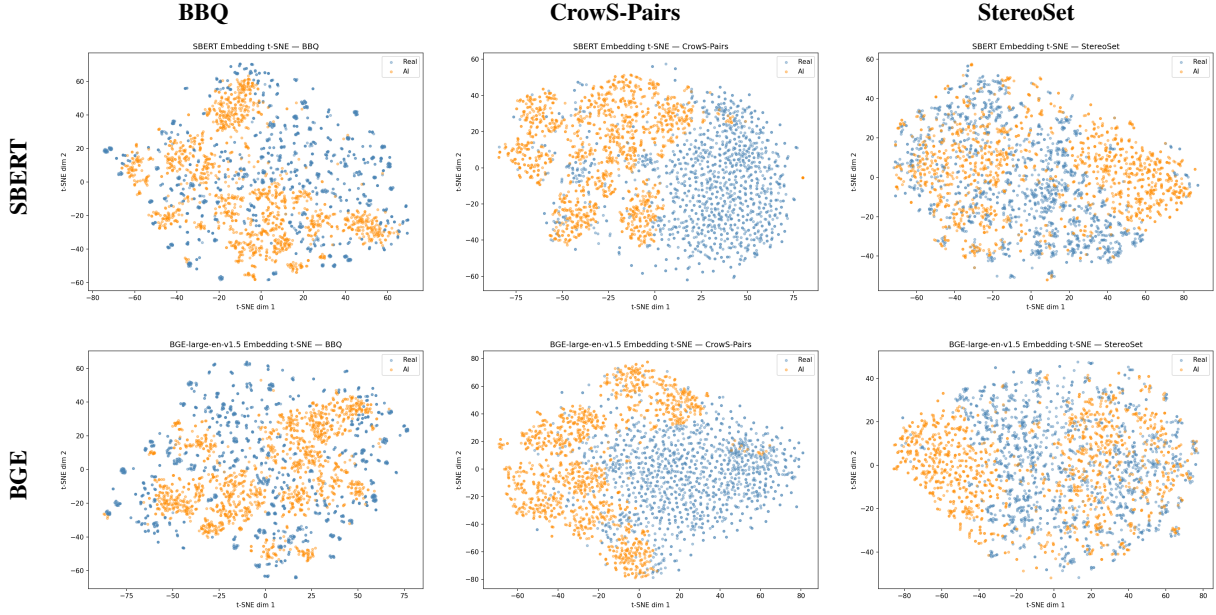


Figure A1: t-SNE projections of sentence embeddings for instances across BBQ, CrowS-Pairs, and StereoSet (blue) and their synthetic counterparts (orange). Top row uses SBERT embeddings; bottom row uses BGE-Large-en-v1.5 embeddings. Synthetic instances are broadly distributed across the embedding space rather than collapsing into narrow clusters, indicating comparable semantic diversity to the original benchmarks.

distributed in embedding space rather than collapsing into narrow semantic clusters. The synthetic data therefore appears to preserve a comparable degree of semantic spread to the original benchmarks. Taken together, these analyses indicate that the synthetic benchmark instances are linguistically well formed, structurally comparable to the original datasets, and semantically diverse enough to support the downstream subsampling experiments reported in the main text.

Finally, we conducted a manual quality audit of the generated instances, reviewing 500 StereoSet instances, 500 BBQ instances, and 200 CrowS-Pairs instances. For each item we checked whether it clearly expressed the intended stereotype/anti-stereotype contrast. The proportion judged clear was 89.8% for StereoSet (449/500), 94.8% for BBQ (474/500), and 88% for CrowS-Pairs (176/200). The remainder consisted largely of culturally specific or niche cases we could not confidently adjudicate, rather than clearly malformed items. These results support the use of the synthetic instances as stereotype evaluation items, but should be read as directional evidence, since a full validation of instance validity is beyond the scope of this paper.

B Multi-Tolerance Evaluation Results with Subsampling on Existing Benchmarks

This section provides the complete full-dataset results for the proposed minimum subsample size prediction method on existing benchmark instances. Figure 1 in the main text illustrates the subsampling behavior at $\varepsilon = 10\%$, while Figures B1-B3 extend that analysis to the remaining tolerance settings. Table B1 complements these figures by reporting the full numerical results across all models, datasets, and tolerances, including the predicted minimum subsample size n^* , the allowed deviation Δ_{pp} , the empirical coverage $\hat{\alpha}$, and the confirmation threshold b_{low} .

Across the full range of tolerances, the same broad patterns remain stable. The deviation histograms are centered close to zero, and the cumulative curves show that the large majority of subsamples remain within the target error bands. CrowS-Pairs and StereoSet exhibit especially tight alignment across models throughout, whereas BBQ shows somewhat broader spread and greater model-specific separation at every tolerance level, although its distributions remain well centered overall. As the tolerance is relaxed from $\varepsilon = 5\%$ to $\varepsilon = 20\%$, the coverage curves rise more steeply and reach near-perfect coverage at lower subsam-

ple ratios, directly reflecting the smaller predicted subsample sizes under looser error criteria. The histograms widen correspondingly as the acceptable deviation band grows, but their mass remains concentrated near zero throughout. Taken together, the figures and Table B1 show that the behavior highlighted in the main text at $\varepsilon = 10\%$ is not specific to that setting, but generalizes reliably across the full range of tolerances considered.

C Multi-Tolerance Evaluation Results with Subsampling on Synthetic Data

This section provides the complete multi-tolerance distributional analysis for synthetic benchmark instances. Figure 2 in the main text illustrates the subsampling behavior at $\varepsilon = 10\%$, and Figures C1-C3 extend that analysis to the remaining tolerance settings. Table C1 complements these figures by reporting the full numerical results across all models, datasets, and tolerances, including the predicted minimum subsample size n^* , the allowed deviation Δ_{pp} , the empirical coverage $\hat{\alpha}$, and the confirmation threshold b_{low} .

Across the full range of tolerances, the same broad behavior persists on synthetic data. Deviation distributions remain centered near zero, and the coverage curves show that the predicted subsample sizes achieve the target coverage in all configurations. CrowS-Pairs and StereoSet again show tight alignment across models, while synthetic BBQ exhibits more gradual coverage growth and requires larger subsample ratios before the target coverage is reached. This pattern is especially visible at stricter tolerances and is consistent with the larger predicted sample sizes for synthetic BBQ in Table C1. A likely reason is that several synthetic BBQ settings have very low p_f values, which make tight relative tolerances more demanding and thus push the required subsample size upward.

As the tolerance increases from $\varepsilon = 5\%$ to $\varepsilon = 20\%$, the coverage curves rise more steeply and reach high coverage at progressively smaller subsample ratios across all three benchmarks. This mirrors the same tolerance-driven efficiency gain observed on existing benchmarks. At the loosest settings, the required subset sizes can become very small, which in turn produces a more discrete set of achievable bias percentages. The multi-modal shapes visible in some 15% and 20% histograms are therefore best understood as a natural consequence of low subset counts, rather than as evi-

dence of unstable evaluation behavior.

Taken together, these figures show that the transfer from existing benchmarks to synthetic benchmark construction is not limited to a single tolerance setting. The same qualitative trends remain visible across all four tolerance levels, providing further support for the robustness of the proposed minimum subsample size prediction method on synthetic data.

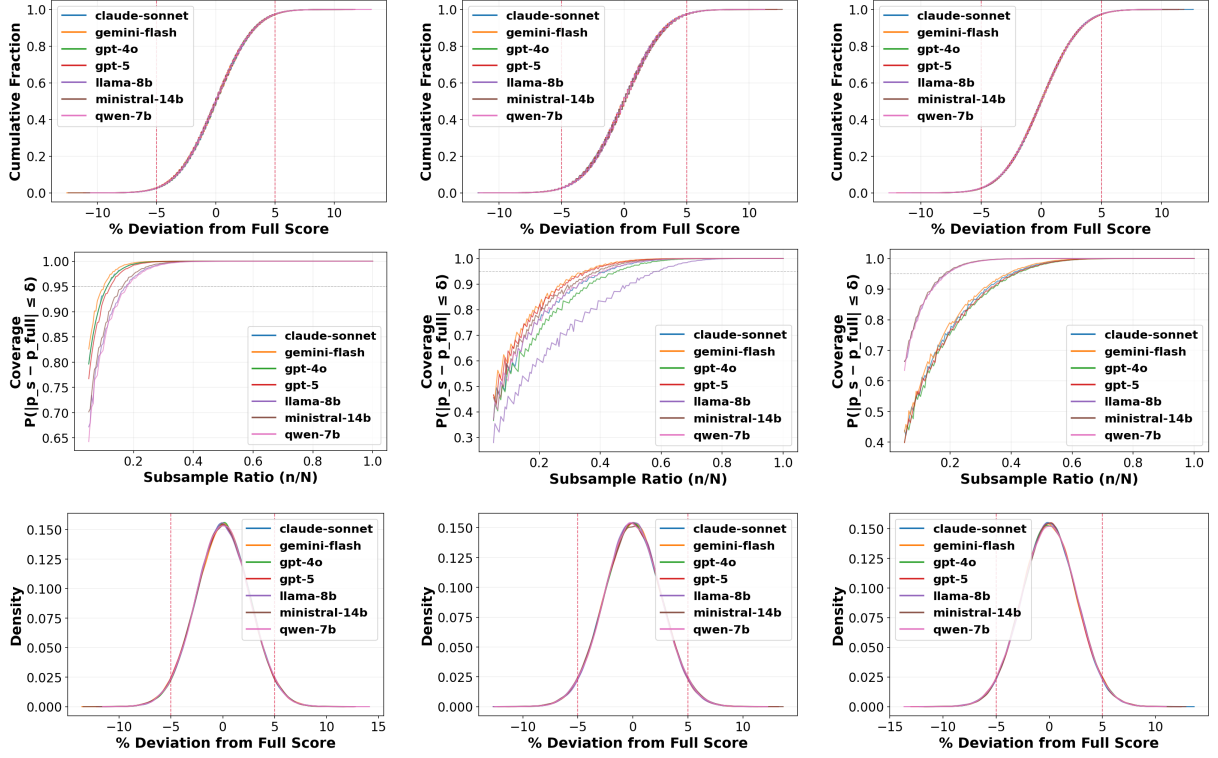


Figure B1: Subsampling behavior on StereoSet, CrowS-Pairs, and BBQ (left to right) at $\varepsilon = 5\%$. Top: empirical cumulative distribution of percent deviation from the full-dataset score. Middle: empirical coverage versus subsample ratio. Bottom: KDE of percent deviation from the full-dataset score.

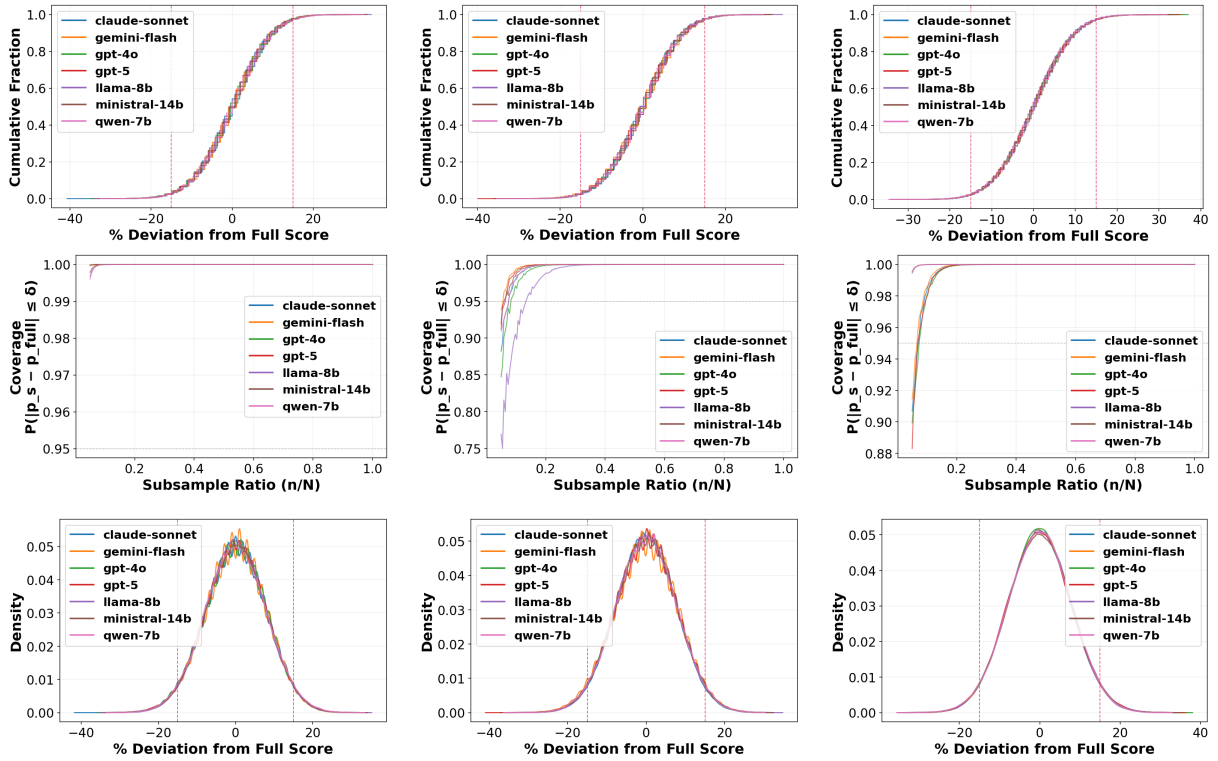


Figure B2: Subsampling behavior on StereoSet, CrowS-Pairs, and BBQ (left to right) at $\varepsilon = 15\%$. Top: empirical cumulative distribution of percent deviation from the full-dataset score. Middle: empirical coverage versus subsample ratio. Bottom: KDE of percent deviation from the full-dataset score.

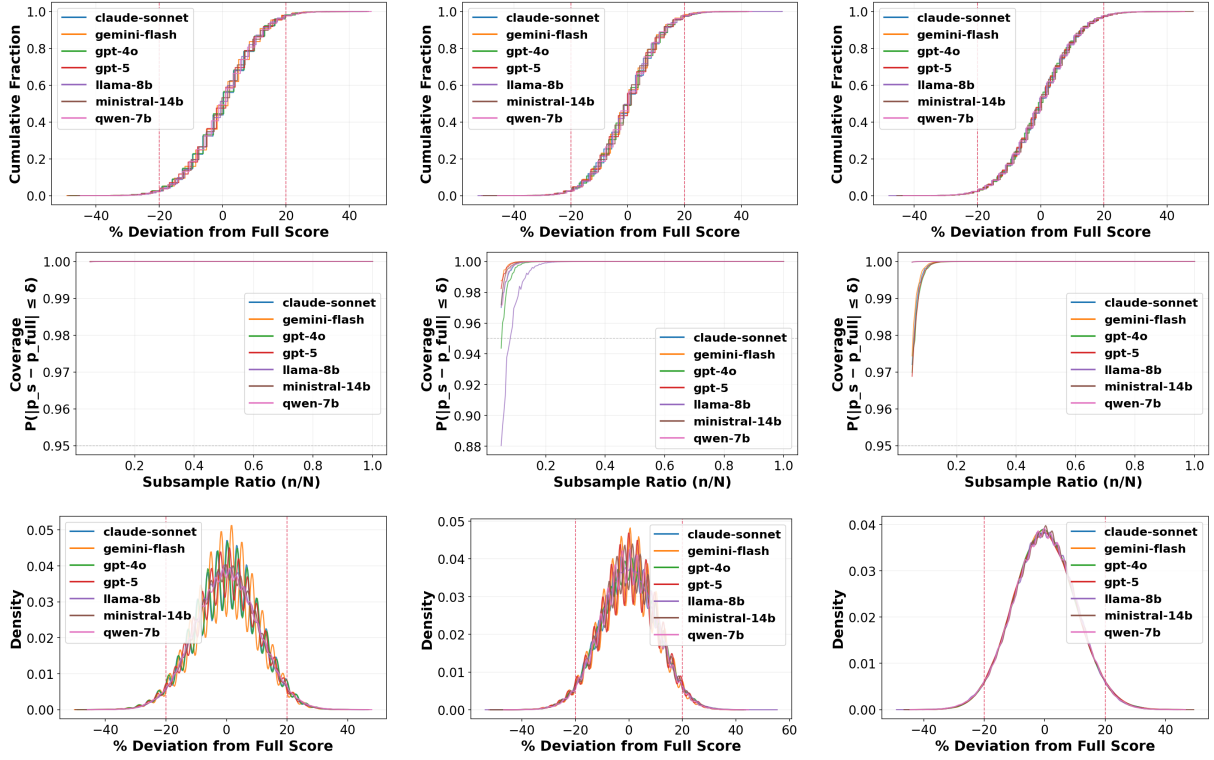


Figure B3: Subsampling behavior on StereoSet, CrowS-Pairs, and BBQ (left to right) at $\varepsilon = 20\%$. Top: empirical cumulative distribution of percent deviation from the full-dataset score. Middle: empirical coverage versus subsample ratio. Bottom: KDE of percent deviation from the full-dataset score.

ε	p_f	Claude Sonnet			Gemini Flash			GPT-4o			GPT-5			LLaMa-8B			Ministral-14B			Qwen-7B		
		BBQ	CrowS	St.Set	BBQ	CrowS	St.Set	BBQ	CrowS	St.Set	BBQ	CrowS	St.Set	BBQ	CrowS	St.Set	BBQ	CrowS	St.Set	BBQ	CrowS	St.Set
5%	n^* (%)	2628 (40%)	615 (41%)	703 (11%)	2526 (38%)	524 (35%)	615 (10%)	2725 (41%)	676 (45%)	706 (11%)	2663 (41%)	531 (36%)	762 (12%)	1221 (19%)	882 (58%)	1095 (17%)	1200 (18%)	570 (38%)	1012 (16%)	1226 (19%)	612 (41%)	1125 (18%)
	Δp_p	1.30	2.95	3.29	1.35	3.28	3.43	1.23	2.76	3.29	1.26	3.23	3.18	2.51	2.11	2.67	2.55	3.08	2.78	2.50	2.97	2.62
	$\hat{\alpha}$	0.953	0.951	0.953	0.951	0.954	0.949	0.951	0.952	0.951	0.951	0.953	0.951	0.952	0.956	0.950	0.953	0.951	0.950	0.951	0.951	0.952
	b_{low}	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949
10%	n^* (%)	930 (14%)	222 (15%)	185 (3%)	871 (13%)	169 (11%)	163 (3%)	978 (15%)	256 (17%)	184 (3%)	957 (15%)	182 (12%)	208 (3%)	352 (5%)	381 (25%)	312 (5%)	348 (5%)	198 (13%)	280 (4%)	355 (5%)	211 (14%)	317 (5%)
	Δp_p	2.59	5.90	6.58	2.70	6.57	6.86	2.46	5.51	6.57	2.52	6.46	6.35	5.02	4.22	5.34	5.10	6.17	5.55	5.01	5.95	5.25
	$\hat{\alpha}$	0.952	0.954	0.950	0.951	0.952	0.951	0.951	0.955	0.951	0.951	0.953	0.951	0.952	0.953	0.951	0.953	0.951	0.951	0.950	0.951	0.951
	b_{low}	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949
15%	n^* (%)	440 (7%)	111 (7%)	82 (1%)	422 (6%)	77 (5%)	75 (1%)	477 (7%)	125 (8%)	84 (1%)	453 (7%)	85 (6%)	92 (1%)	163 (2%)	200 (13%)	145 (2%)	152 (2%)	95 (6%)	127 (2%)	162 (2%)	104 (7%)	141 (2%)
	Δp_p	3.89	8.85	9.88	4.06	9.85	10.30	3.70	8.27	9.86	3.78	9.69	9.53	7.53	6.34	8.01	7.64	9.25	8.33	7.51	8.92	7.87
	$\hat{\alpha}$	0.951	0.955	0.954	0.954	0.952	0.955	0.952	0.952	0.950	0.951	0.953	0.951	0.953	0.955	0.956	0.952	0.951	0.954	0.952	0.951	0.951
	b_{low}	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949
20%	n^* (%)	255 (4%)	61 (4%)	47 (1%)	244 (4%)	47 (3%)	43 (1%)	278 (4%)	70 (5%)	47 (1%)	262 (4%)	48 (3%)	53 (1%)	92 (1%)	115 (8%)	83 (1%)	90 (1%)	56 (4%)	74 (1%)	92 (1%)	62 (4%)	83 (1%)
	Δp_p	5.18	11.81	13.17	5.41	13.13	13.73	4.93	11.02	13.15	5.04	12.92	12.71	10.03	8.45	10.68	10.19	12.33	11.11	10.01	11.90	10.50
	$\hat{\alpha}$	0.950	0.954	0.956	0.953	0.959	0.954	0.953	0.952	0.955	0.951	0.955	0.956	0.954	0.951	0.954	0.956	0.950	0.956	0.954	0.954	0.954
	b_{low}	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949

Table B1: Minimum subsample prediction on existing benchmark instances. Each cell block reports, for each model $\times$ dataset $\times$ relative tolerance ε , the required subsample size n^* (% of full set N), the allowed deviation from the full-set score in percentage points ($\Delta p_p = \varepsilon \cdot p_f$), the empirical coverage $\hat{\alpha}$, and the confirmation lower bound b_{low} . A prediction is confirmed when $\hat{\alpha} \geq b_{low}$. The full-set stereotype score p_f is shown under each dataset.

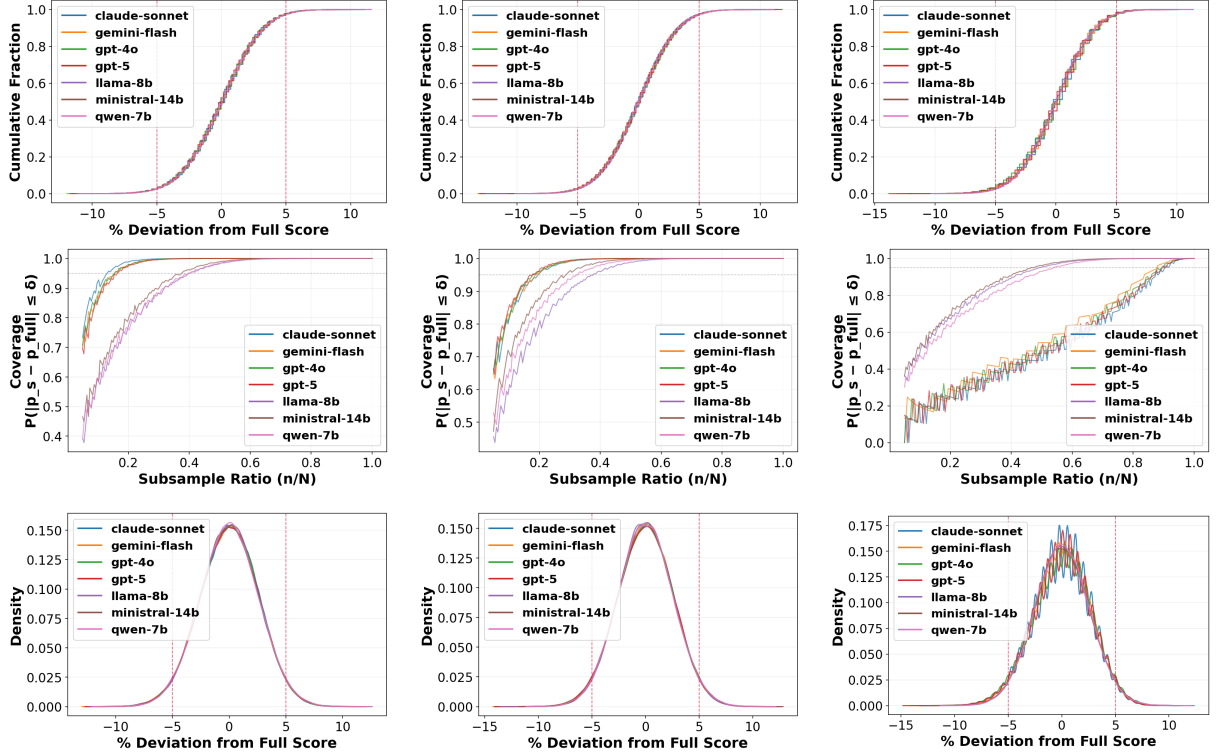


Figure C1: Subsampling behavior on synthetic data generated in the formats of StereoSet, CrowS-Pairs, and BBQ (left to right) at $\varepsilon = 5\%$. Top: empirical cumulative distribution of percent deviation from the full-dataset score. Middle: empirical coverage versus subsample ratio. Bottom: KDE of percent deviation from the full-dataset score.

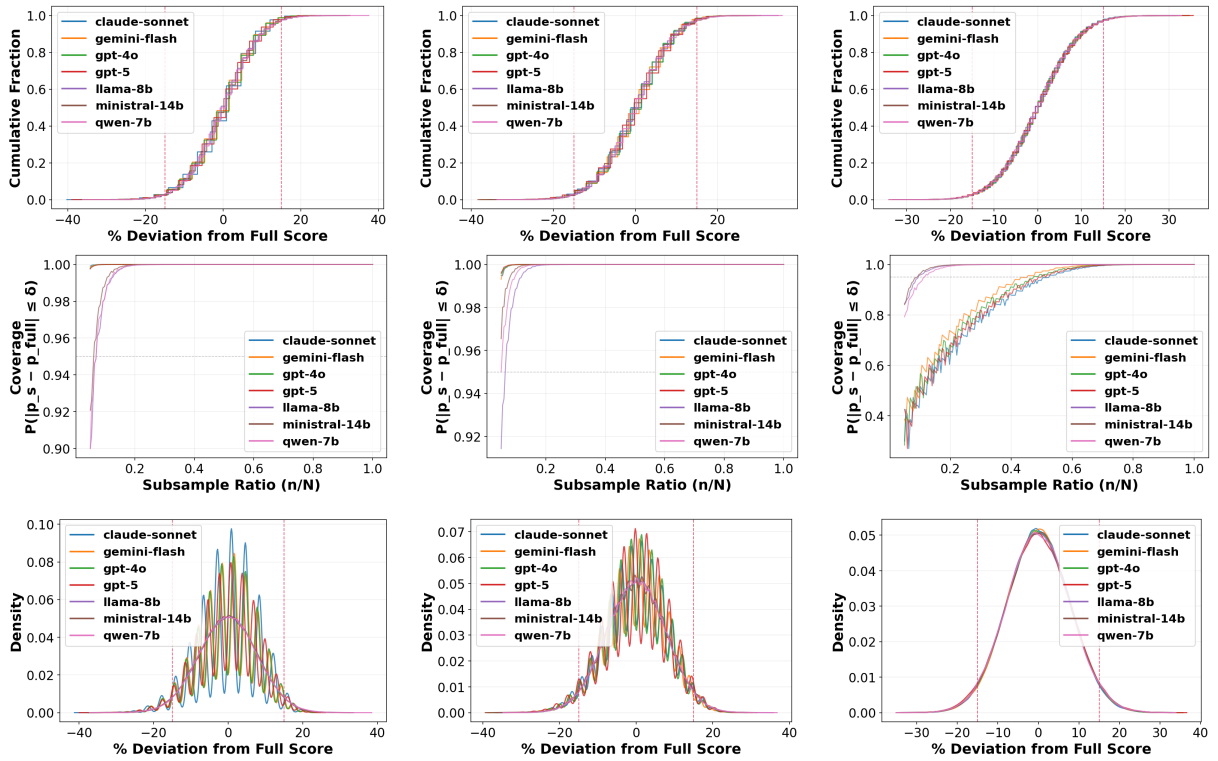


Figure C2: Subsampling behavior on synthetic data generated in the formats of StereoSet, CrowS-Pairs, and BBQ (left to right) at $\varepsilon = 15\%$. Top: empirical cumulative distribution of percent deviation from the full-dataset score. Middle: empirical coverage versus subsample ratio. Bottom: KDE of percent deviation from the full-dataset score.

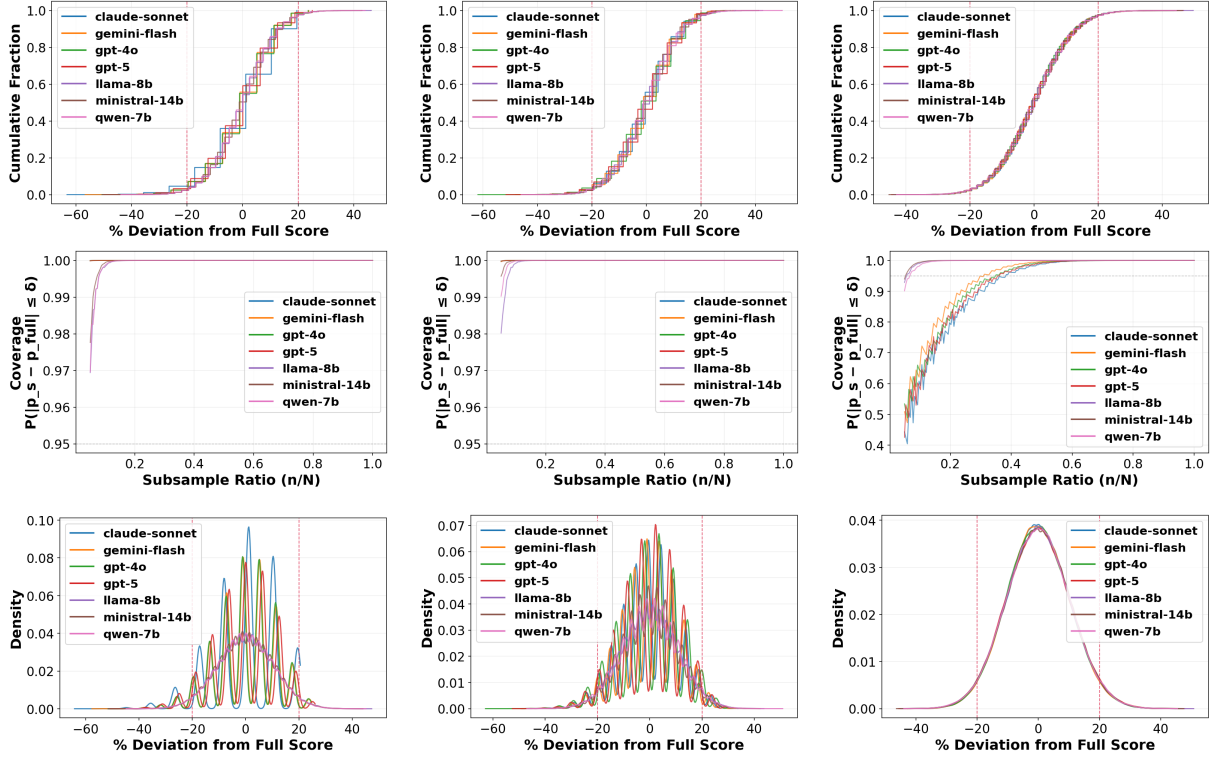


Figure C3: Subsampling behavior on synthetic data generated in the formats of StereoSet, CrowS-Pairs, and BBQ (left to right) at $\varepsilon = 20\%$. Top: empirical cumulative distribution of percent deviation from the full-dataset score. Middle: empirical coverage versus subsample ratio. Bottom: KDE of percent deviation from the full-dataset score.

ε	p_f	Claude Sonnet			Gemini Flash			GPT-4o			GPT-5			LLaMa-8B			Ministral-14B			Qwen-7B		
		BBQ	CrowS	St.Set	BBQ	CrowS	St.Set	BBQ	CrowS	St.Set	BBQ	CrowS	St.Set	BBQ	CrowS	St.Set	BBQ	CrowS	St.Set	BBQ	CrowS	St.Set
5%	n^* (%)	2228 (90%)	402 (19%)	254 (13%)	2163 (88%)	386 (18%)	301 (15%)	2176 (88%)	370 (18%)	303 (15%)	2209 (89%)	352 (17%)	316 (16%)	1209 (49%)	787 (38%)	796 (40%)	1144 (46%)	587 (28%)	731 (37%)	1329 (54%)	666 (32%)	789 (39%)
	Δp_p	0.30	3.77	4.18	0.40	3.79	4.04	0.35	3.82	4.04	0.33	3.87	4.00	1.95	2.75	2.66	2.07	3.25	2.83	1.74	3.03	2.69
	$\hat{\alpha}$	0.955	0.954	0.955	0.958	0.953	0.954	0.951	0.950	0.954	0.954	0.950	0.954	0.953	0.955	0.952	0.952	0.954	0.951	0.954	0.951	0.951
	b_{low}	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949
10%	n^* (%)	1750 (71%)	115 (5%)	69 (3%)	1583 (64%)	108 (5%)	85 (4%)	1670 (68%)	106 (5%)	81 (4%)	1703 (69%)	99 (5%)	82 (4%)	478 (19%)	266 (13%)	287 (14%)	441 (18%)	178 (8%)	251 (13%)	551 (22%)	216 (10%)	272 (14%)
	Δp_p	0.60	7.53	8.37	0.79	7.58	8.08	0.69	7.65	8.09	0.65	7.74	7.99	3.90	5.51	5.33	4.13	6.49	5.66	3.48	6.07	5.38
	$\hat{\alpha}$	0.960	0.954	0.955	0.956	0.950	0.953	0.957	0.954	0.958	0.958	0.952	0.953	0.953	0.951	0.953	0.953	0.951	0.952	0.953	0.953	0.949
	b_{low}	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949
15%	n^* (%)	1284 (52%)	51 (2%)	32 (2%)	1097 (44%)	51 (2%)	39 (2%)	1170 (47%)	49 (2%)	39 (2%)	1189 (48%)	44 (2%)	41 (2%)	232 (9%)	128 (6%)	134 (7%)	213 (9%)	85 (4%)	120 (6%)	270 (11%)	102 (5%)	131 (7%)
	Δp_p	0.90	11.30	12.55	1.19	11.37	12.11	1.04	11.47	12.13	0.98	11.62	11.99	5.85	8.26	7.99	6.20	9.74	8.48	5.22	9.10	8.08
	$\hat{\alpha}$	0.958	0.952	0.949	0.956	0.955	0.963	0.954	0.961	0.963	0.949	0.958	0.954	0.954	0.957	0.952	0.951	0.952	0.955	0.951	0.951	0.955
	b_{low}	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949
20%	n^* (%)	927 (38%)	28 (1%)	13 (1%)	768 (31%)	28 (1%)	20 (1%)	844 (34%)	24 (1%)	20 (1%)	858 (35%)	24 (1%)	20 (1%)	137 (6%)	70 (3%)	77 (4%)	124 (5%)	48 (2%)	68 (3%)	161 (7%)	55 (3%)	76 (4%)
	Δp_p	1.21	15.07	16.73	1.59	15.16	16.15	1.38	15.30	16.18	1.30	15.49	15.99	7.80	11.02	10.65	8.26	12.98	11.31	6.97	12.13	10.77
	$\hat{\alpha}$	0.956	0.956	0.954	0.956	0.957	0.961	0.956	0.951	0.961	0.952	0.956	0.957	0.953	0.949	0.954	0.952	0.955	0.956	0.951	0.950	0.955
	b_{low}	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949	0.949

Table C1: Minimum subsample prediction on synthetic instances. Each cell block reports, for each model $\times$ dataset $\times$ relative tolerance ε , the required subsample size n^* (% of full set N), the allowed deviation from the full-set score in percentage points ($\Delta p_p = \varepsilon \cdot p_f$), the empirical coverage $\hat{\alpha}$, and the confirmation lower bound b_{low} . A prediction is confirmed when $\hat{\alpha} \geq b_{low}$. The full-set stereotype score p_f is shown under each dataset.